# METHOD AND SYSTEM TO COMPENSATE FOR THE EFFECTS OF PACKET DELAYS ON SPEECH QUALITY IN A VOICE-OVER IP SYSTEM

## BACKGROUND OF THE INVENTION

### Field of the Invention.

**[0001]** The present invention generally relates to voice-over Internet Protocol (VoIP) systems and more particularly to a method and system for compensating for the effects of packet delays on perceived speech quality in a VoIP system.

### Description of Related Art

**[0002]** Voice Over IP (VoIP) is short for Voice over Internet Protocol. It includes any technology that enables voice telephony over IP Networks. If VoIP is used in a private setting such as in an Intranet, or over a Wide Area Network (WAN), it is generally referred to as VoIP. If VoIP is transported over the public Internet, VoIP is generally referred to as Internet telephony.

**[0003]** Speech signals are transmitted over packet networks in a VoIP system. Due to real-time constraints on the delivery of speech signals, the perceived speech quality is susceptible to delayed arrival of packets. In a packet network, packets may arrive asynchronously, i.e. the exact arrival time of a packet can be significantly different from a nominal arrival time. Also, each packet may be subject to differing amounts of packet delay in the network. This variation in packet delay is called jitter.

As a result of jitter, packets may arrive out of sequence at an end-point, wherein an end-point is a client in a server-client network, for example.

**[0004]** Packets arriving at the end-point are queued in a jitter buffer where they are placed back in sequence, if required. In a VoIP system, a jitter buffer is typically required wherever a speech packet stream is terminated. The purpose of the jitter buffer is to absorb jitter. Since packets arrive asynchronously, they are placed into the jitter buffer asynchronously, but the packets are removed from the jitter buffer synchronously, i.e. at fixed time intervals. The jitter buffer is typically permitted to grow to a certain size before packets are removed for processing. This size of the jitter buffer is referred to as the nominal length of the jitter buffer. Since packets are arriving at random time intervals, the instantaneous number of packets in the jitter buffer may be different from the nominal length. This different length may be referred to as the instantaneous length or jitter buffer length. The maximum length to which a jitter buffer may grow is referred to as the maximum size of the jitter buffer. The sequence of the incoming packets is monitored and if a missing (or out of sequence) packet is detected, a placeholder or NULL packet is created to take the out of sequence packet's place. If the missing packet arrives at a subsequent time, it may replace this NULL packet. If the missing packet has not arrived by the time it is supposed to be removed from the jitter buffer, a packet loss is declared by the VoIP system. These packet losses can result in significant degradation in the perceived speech quality at a receiver that is receiving these packets.

## SUMMARY OF THE INVENTION

**[0005]** The present invention provides a method and system to compensate for packet delay in a Voice over Internet Protocol (VoIP) system. The system includes a jitter buffer. The jitter buffer may be embodied as a queue for receiving speech packets in the VoIP system. The system includes a variable-speed playback device for adjusting the playback speed of the received speech packets, and a jitter buffer manager for detecting packet jitter, and for sending commands to the playback device to adjust playback speed based on the detection.

**[0006]** In the system, the signal is played back at a nominal speed when there are no out of sequence packets. The playback speed is decreased when an out of sequence packet is detected. This tends to increase the length of the jitter buffer. When a delayed packet arrives, the playback speed is increased in order to restore the jitter buffer length to its nominal value.

**[0007]** The present invention is based upon an observation that human hearing is not very sensitive to small variations in the speed at which a speech signal is played back, so long as the pitch (i.e., fundamental frequency of the speech signal) is maintained. Variable Speed Playback (VSP) of speech signals is used to adjust the jitter buffer length, but the average jitter buffer length is kept relatively small. The jitter buffer length adjustment is transparent to the user. Hence, the perceived end-to-end delay is maintained as small as in a conventional VoIP system, which has a smaller jitter buffer length, while significantly improving the perceived speech quality.

## BRIEF DESCRIPTION OF THE DRAWINGS

**[0008]** The present invention will become more fully understood from the detailed description given below and the accompanying drawings, wherein like elements are represented by like reference numerals, which are given by way of illustration only and thus are not limitative of the present invention and wherein:

**[0009]** Fig. 1 illustrates a functional block diagram of the system in accordance with the invention;

**[0010]** Fig. 2 illustrates attributes of jitter buffer 100 in accordance with the invention;

**[0011]** Fig.3 illustrates a state-machine representation of operations performed on the playback device 200 in accordance with the invention;

**[0012]** Fig. 4 illustrates a flow-chart of the operations performed by the jitter buffer manager 300 in accordance with the invention; and

**[0013]** Fig. 5 illustrates an exemplary application performed by the system in accordance with the present invention.

## DETAILED DESCRIPTION

**[0014]** The effects of packet delay may, in theory, be compensated by increasing the length of the jitter buffer. However, it is not possible to increase the jitter buffer length arbitrarily in a real-time environment. It is desirable to maintain the jitter buffer length small when packet jitter is low in order to minimize latency on a VoIP system. Accordingly, the present invention is based upon the observation that human hearing is not very sensitive to small variations in the speed at which a speech signal is played back, as long as the pitch (such as the frequency of vibration of the vocal chords) is maintained. Due to Variable Speed Playback (VSP) of speech in the system, the jitter buffer length may vary. However, the jitter buffer manager attempts to maintain jitter buffer length around the nominal value. Any jitter buffer length adjustment is transparent to the user and hence, the perceived end-to-end delay is maintained as small as in a conventional VoIP system, while significantly improving the perceived speech quality.

**[0015]** In the system, the signal is played back at a nominal speed when there are no out of sequence packets, i.e. when no NULL packets have been placed in the jitter buffer. The playback speed is decreased when an out of sequence packet is detected. This tends to increase the jitter buffer length. When the out of sequence packet arrives or is declared to be a lost packet by the system, the playback speed is increased in an attempt to restore the jitter buffer length to its nominal value.

**[0016]** Fig. 1 illustrates a functional block diagram of the system in accordance with the invention. As shown in Fig. 1, the system includes a jitter buffer 100, variable speed playback (VSP) device 200 and jitter buffer manager 300. Incoming speech packets are received at jitter buffer 100. Jitter buffer manager 300 monitors jitter buffer 100 and controls playback speed of VSP device 200 via path 110. Control of playback speed is based on the status of the jitter buffer 100. VSP device 200 receives speech packets 105 from jitter buffer 100, and plays back speech signals 115 representing speech packets 105. These speech signals 115 are transmitted to external receivers, external mediums, and/or other components in the VoIP system

**[0017]** The jitter buffer 100 may be a queue for receiving incoming speech packets 105. The jitter buffer 100 is used to absorb jitter that is inherent in a VoIP system. As noted above, speech packets 105 may arrive out of sequence at the jitter buffer 100, and are placed back in the correct sequence at jitter buffer 100. When playback speed is decreased, the length of the jitter buffer 100 tends to increase. The jitter buffer length tends to decrease when the playback speed is increased.

**[0018]** The VSP device 200 may be embodied as any known or developing variable speed playback device known in the art. The VSP device 200 varies the playback speed of a speech signal by a specified amount using a method such as time-domain harmonic scaling or time-scale modification. A more detailed description of time-domain harmonic scaling can be found in an article by David Malah, entitled TIME-DOMAIN ALGORITHMS FOR HARMONIC BANDWIDTH REDUCTION AND TIME SCALING OF SPEECH SIGNALS, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-27, N.2, April 1979, the contents of which are incorporated by reference herein in their entirety. Additionally, time-scale modification is described in the article by Michael R. Portnoff, entitled TIME-SCALE MODIFICATION OF SPEECH BASED ON SHORT-TIME FOURIER ANALYSIS, IEEE Transactions on Acoustics, Speech, and Signal Processing, vol. ASSP-29, No.3, June 1981, the contents of which are incorporated by reference herein in their entirety.

**[0019]** The buffer manager 300 may be embodied as an algorithm, application or process run by, or implemented in, a general purpose processor or microprocessor, a specialized processor such as a digital signal processor (DSP) or an application-specific integrated circuit (ASIC). The jitter buffer manager 300 and jitter buffer 100 may be further provided as a single component or circuit, or provided as separate components on substrate or surface of a chip, circuit or component

**[0020]** Fig. 2 illustrates attributes of jitter buffer 100 in accordance with the invention. As seen in Fig. 2, the Nominal Length is the expected steady state length of the jitter buffer 100. The High Water Mark is a maximum length to which jitter buffer 100 is permitted to grow. The difference between the Nominal Length and the High Water Mark is termed T_high_water in Fig. 2. The actual maximum jitter buffer size can be higher than the High Water Mark if, for instance, the jitter buffer 100 has already reached the High Water Mark and a burst of packets arrives before the system has had time to reduce the jitter buffer length.

**[0021]** When an out of sequence packet is detected in the jitter buffer 100, a placeholder NULL packet is created in its place. The parameter T_early is defined as follows. When a NULL packet location 125 is more than T_early packets from the head of the jitter buffer 100 (i.e., the earliest packet in the jitter buffer 100) playback speed control can be disabled by jitter buffer manager 300. This feature may be used to prevent rapid changes in playback speed when the jitter buffer length is close to the High Water Mark.

**[0022]** Fig. 3 illustrates a state-machine representation of the control operations performed on the VSP device 200 by the jitter buffer manager 300. Each state corresponds to a playback speed. There are three possible speeds: Nominal, High and Low. The default speed is the nominal playback speed. As long as there are no NULL packets in the jitter buffer 100, or so long as all NULL packets are at least T_early packets away from the head of the jitter buffer 100, the playback speed remains unchanged.

If the jitter buffer manager 300 detects a NULL packet within T_early packets of the head of the jitter buffer 100, the jitter buffer manager 300 sets the playback speed of the VSP device 200 to Low.

**[0023]** While the playback speed is Low, if a NULL packet is filled up by the delayed arrival of the out of sequence packet, or if the out of sequence packet is declared to be lost, the jitter buffer manager 300 sets the playback speed of the VSP device 200 to High in order to restore the length of jitter buffer 100 to its nominal length. The jitter buffer manager 300 also sets the playback speed of the VSP device 200 to High if the length of jitter buffer 100 reaches a High Water Mark. This is done to prevent the jitter buffer 100 from growing indefinitely in size.

**[0024]** The jitter buffer manager 300 maintains the playback speed of the VSP device 200 High as long as the length of jitter buffer 100 is larger than its nominal length. When the length of the jitter buffer 100 returns or decreases to its nominal length, the jitter buffer manager 300 sets the playback speed of the VSP device 200 to Nominal.

**[0025]** Fig. 4 illustrates operations performed by the jitter buffer manager 300. Initially, the jitter buffer manager 300 checks the current state (i.e. the current speed setting) of the VSP device 200. (Steps S1 and S2). If the current speed setting is Nominal, the jitter buffer manager 300 computes a difference D between the head of the jitter buffer 100 and the location of the earliest NULL packet (Step S9). If D is greater than T_early (Step S10), there is no state change, i.e. the playback speed remains Nominal (Step S11). If D is less than or equal to T_early, the jitter buffer manager 300 changes the playback speed of the VSP device 200 to Low (Step S12).

**[0026]** If the current speed setting is High, the current length of jitter buffer 100 is checked against the nominal length (Step S3). If the current length of jitter buffer 100 exceeds the nominal length, there is no state change, i.e. the playback speed is maintained High in VSP device 200 (Step S8). Otherwise, the jitter buffer manager 300 changes the playback speed of the VSP device 200 to Nominal (i.e. the current length of jitter buffer 100 has become equal to the nominal length) (Step S11).

**[0027]** If the current speed setting is Low, jitter buffer manager 300 computes a difference D between the head of the jitter buffer 100 and the location of the earliest NULL packet. (Step S4). If D is determined to exceed T_early (Step S5), the playback speed is set to High (Step S8). Otherwise, the current length of the jitter buffer 100 is checked against the High Water Mark (Step S6). If it has reached the High Water Mark, the playback speed is also set to High (Step S8). Otherwise, there is no state change, i.e. the playback speed is kept Low (Step S7). Accordingly, by intelligently varying the playback speed according to the above algorithm, additional time is allocated in the system for an out of sequence packet to arrive.

**[0028]** If silence suppression is used in the system, received silence intervals may also be used to adjust the length of jitter buffer 100 by controlling the playback speed. For instance, if the length of jitter buffer 100 is exceeding the Nominal Length when a silence period is detected, the buffer manager 300 controls the VSP device 200 so that playback may occur at the Nominal speed instead of the faster (High) speed. This effectively compresses the silence interval. When a new burst of speech packets arrives, the jitter buffer 100 is empty and the jitter buffer manager 300 instructs the VSP device 200 to start playing out the new speech packets at the Nominal speed.

## Exemplary Applications

**[0029]** The system and method of the present invention have several possible applications. As noted earlier, a jitter buffer is required wherever a speech packet stream is terminated in a VoIP system, i.e., at an end-point. The system and method of the present invention may be utilized to alleviate the effects of packet delays at all such end-points. A few exemplary applications are briefly described.

**[0030]** The system and method may be used in a "Conferencing over IP" application. A conference is when a number of end-points or clients simultaneously participate in a conversation. Briefly, each client can listen to the speech transmitted by every other client in the conference. The mechanism that enables this multi-party conferencing is called a conferencing bridge.

**[0031]** Fig. 5 illustrates an exemplary conferencing over IP application using the system and method of the invention. In Fig. 5, there are a number of clients 500A-N participating in the conference. Each client 500 sends out a speech packet stream 501 that is terminated at the Conference Bridge 600. Each of these speech packet streams 501 has to pass through a jitter buffer 603 before the Conference Bridge 600 can process the speech signals 602. The Conference Bridge 600 mixes the speech signals 602 from all the clients 500A-N using mixing algorithms that are well-known in the art.

**[0032]** Conference Bridge 600 produces an outgoing speech packet stream 601 for each client 500. The client 500, now as an end-point, terminates this stream 601 so it has to pass through jitter buffer 503, to terminate at each client 500 as a speech signal 502. The system of the present invention would operate at the location of each jitter buffer 503 and 603, i.e. at the conference bridge 600 and at clients 500A-N.

**[0033]** Another application is in an IP-based Voice Messaging environment. Voice messages sent over IP are recorded and voice messages are played back over IP by this application. The system could reduce distortion due to packet delays while recording, as well as while performing playback functions.

**[0034]** A further application is in streaming a voice broadcast over internet. Streaming a voice broadcast over internet would include live web-cast of news, sports event coverage, company-wide speeches etc. A typical user application that terminates a streaming voice broadcast buffers up voice packets before beginning to play out the signal. In the event of network congestion, frequent re-buffering is required during which a number of packets are lost and the voice broadcast is interrupted. The system described in this invention could reduce the number of lost packets and therefore reduce the need for re-buffering.

**[0035]** Yet a further application of the method and system is within an Automatic Speech Recognition (ASR) over IP system. In a number of voice-enabled services or applications, voice recognition is performed on voice coming over IP. In this system, the end-user of the voice packet is not a human being, but an application. Packet delays result in packet losses, which may result in discontinuities and distortions in the speech signal. These distortions result in a degradation of the performance of ASR in the system. The system of the invention could be used to reduce the packet loss and hence the speech signal distortion. This may result in an improvement in the performance of the ASR over IP system.

**[0036]** Additionally, the system and method could be applied to E-mail reading over the internet, a feature which converts a text message into speech using a technology such as Text-To-Speech Synthesis (TTS). The TTS feature transmits the synthesized speech over IP. The system and method of the present invention would preferably operate at the end points, so as to reduce degradation of speech quality due to packet delay. The invention being thus described, it will be obvious that the same may be varied in many ways. For example, the functional blocks or steps in Figs. 1-5 may be implemented in hardware and/or software. The hardware/software implementations may include a combination of processor(s) and article(s) of manufacture. The article(s) of manufacture may further include storage media and executable computer program(s). The executable computer program(s) may include the instructions to perform the described operations. The computer executable program(s) may also be provided as part of externally supplied propagated signal(s). Such variations are not to be regarded as departure from the spirit and scope of the invention, and all such modifications as would be obvious to one skilled in the art are intended to be included within the scope of the following claims.